Enriching e-learning metadata through digital library usage analysis

Núria Ferran, *Information and Communication Sciences Studies, Universitat Oberta de Catalunya, Barcelona, Spain*
Jaume Casadesús, *IRTA, Lleida, Spain*
Monika Krakowska, *Institute of Information and Library Science, Jagiellonian University, Krakow, Poland*
Julià Minguillón, *Department of Computer Science, Multimedia and Telecommunication, Universitat Oberta de Catalunya, Barcelona, Spain*

{nferranf, jminguillona}@uoc.edu
jaume.casadesus@irta.es
krakowska@inib.uj.edu.pl

Keywords: digital libraries, recommender systems, learning objects, metadata, personalization, e-learning, usage analysis

Purpose
In this paper we propose an evaluation framework for analyzing learning objects usage, with the aim of extracting useful information for improving the quality of the metadata used to describe the learning objects, but also for personalization purposes, including user models and adaptive itineraries.

Methodology
We present experimental results from the log usage analysis during one academic semester of two different subjects, 350 students. The experiment looks into raw server log data generated from the interactions of the students with the classroom learning objects, in order to find relevant information that can be used to improve the metadata used for describing both the learning objects and the learning process.

Findings
Preliminary studies have been carried out in order to obtain an initial picture of the interactions between learners and the virtual campus, including both services and resources usage. These studies try to establish relationships between user profiles and their information and navigational behavior in the virtual campus, with the aim of promoting personalization and improving the understanding of what learning in virtual environments means.

Research limitations
During the formal learning process, students use learning resources from the virtual classroom provided by the academic library, but they also search for information outside the virtual campus. Not all of these usage data are considered in the model we propose. Further research needs to be done in order to get a complete view of the information search behavior of students for improving the users' profile and creating better personalized services.

Practical implications
In this paper we suggest how a selection of fields used in the LOM standard could be used for enriching the description of learning objects, automatically in some cases, from the learning objects usage performed by an academic community.

Originality
Ever since the beginnings of libraries, they have been a “quiet storage place”. With the development of digital libraries, they become a meeting place where explicit and implicit recommendations about information sources can be shared among users. Social and learning process interactions, therefore, can be considered another knowledge source.

Article type: Research paper

1. Introduction

The intensive use of Information and Communication Technologies such as the Internet increases the possibilities for both content searching and delivery. This new paradigm has completely changed the vision in the distance education field. For example, web-based learning scenarios are becoming a common tool for both face to face and distance educational institutions. E-learning is one of the most promising and growing issues in the information society nowadays, mainly because the growth of the Internet is bringing online education to people in corporations, institutes of higher education, the government and other sectors (Rosenberg, 2002). The growing need of continuous education and the inclusion of new multimedia technologies become crucial factors for this expansion. This fact is supported by two important issues: first, the appearance of new e-learning standards for describing complex learning scenarios, such as IMS-LD (IMS, 2003) and SCORM 2004 (ADL, 2004), and second, the new trends in education defined by what is known as the Bologna process (Bologna Declaration, 1999), where learners become the center of any educational experience, and all the activities, resources and scheduling are arranged according to each learner.

Nowadays, it becomes necessary to shift from heavily content-based courses to other formative actions where the activity is the key concept. Today, activities and the competences developed by such activities are becoming the focus of any formative action. It is also important to promote the formal acknowledgement of skills, knowledge and competences gained through work experience, non-formal training and life experience, for prior learning recognition purposes. This setup promotes what is known as a lifelong learning scenario, where learners continuously improve their competences and knowledge by selecting the best formative activities according to their preferences, particularities and specific needs. Nevertheless, high quality educational contents become the basic elements of this new learning process, but it is worth remarking that this learning process is user-centered, not content-centered.

In the last few years, personalization has become an important issue for both instructional designers and teachers. The high diversity of user profiles and backgrounds, and the new scenario defined by the Bologna Declaration makes it necessary to personalize the learning process for each learner, according to his or her preferences, particularities, competences, and so. Content personalization allows teachers to describe adaptive itineraries depending on the necessities of each known profile, in order to deliver the appropriate content

for each learner. Personalization is strongly related to user experience and satisfaction, which are supposed to be linked to academic performance and dropout rates, in the case of a virtual learning scenario. Furthermore, several studies have tried to discover the relevant attributes of the service quality in online environments, and in most of them, personalization was considered to be one of these essential attributes (Lee and Lin, 2005; Voss, 2003) among others, such as quality of the content, usability or reliability.

In order to do so, providing adaptive itineraries needs an underlying architecture where contents are highly structured and they can be properly stored, described and accessed in a dynamic framework. Learning object repositories (LOR) become a crucial element of any lifelong learning scenario, as they give support to the learning process. Therefore digital libraries, as providers of heterogeneous educational resources, are also essential. Today, digital libraries are also a target for personalization purposes, especially when they are integrated in a virtual learning environment (Ferran et al., 2005).

In this paper we propose an evaluation framework for rating learning objects usage, with the aim of extracting useful information for improving the quality of the metadata used to describe such learning objects, but also for personalization purposes in adaptive systems (Towle and Halm, 2005). We also propose the use of an ontology for establishing the data collection process and ensuring a high degree of coherence between all the elements in the learning scenario, that is, user profiles, learning objects, interactions and so. Two study cases about the usage of several learning resources in a digital library related to a course in Statistics and a course in Archive Management are used to exemplify the relationships between user profiles, content delivery and personalization issues.

2. Learning objects usage in e-learning environments

In a lifelong learning scenario, learners could follow informal, non-formal and formal processes of learning (Coombs and Ahmed, 1974). In the first two options, the learning cycle can be based on searching the Internet and selecting the most suitable contents. Students in these processes do not need to get involved with any institution, except for mentoring, evaluation and accreditation purposes. On the contrary, learners in a university receive, through its set of instructional designers, teachers, and librarians, pushed content with very few options to decide what, where and when to study.

In both processes, there is the need for describing all the learning objects using the appropriate metadata. Learning online needs appropriate metadata for describing the content in order to retrieve and select it, and online learning needs metadata for describing not only the content in itself but also competences and even user preferences for building adaptive itineraries.

Learning objects can be considered as “any entity, digital or non-digital, which can be used, re-used or referenced during technology supported learning” (LOM, 2000). At an educational level, libraries have traditionally been the

institutions responsible for organizing and providing metadata to these resources. If we focus on a digital library as a collection of educational material, it may be useful to think of it as a repository of learning objects. Basically, a repository stores educational resources and their descriptions for providing access and retrieval for teachers and students, either through the Internet or with locked up access behind passwords within proprietary systems.

Librarians describe the resources of catalogues and other collections trough metadata in order to facilitate efficiently the delivery of information. The resource descriptions enable users to discover and identify existing materials and to evaluate and distinguish between different resources and allow the option to personalize the information presented, via the learner's information profile (Foster-Jones and Beazleigh, 2002). Metadata provide controlled and structured descriptions of resources through searchable access points such as title, author, date, location, description and subject, but can also provide interpretative information on the potential educational application of resources or include described information about the relations between the resources (Friesen, 2002). The distinction of these two types of educational metadata is labeled as authoritative and non-authoritative (Recker and Wiley, 2001).

Unlike authoritative metadata, which are generally managed by librarians, non-authoritative metadata are more likely to be generated by the final users of learning objects, so it is interesting to automatically produce these quality data by tagging the navigational actions performed by particular user profiles with learning objects, in order to meet students and teachers expectations when using a LOR. As regards personalization purposes, LORs provide different services to users, from the searching/browsing possibilities to personal services such as a system that keeps track of user interests based on which educational resources he or she searches and downloads. For instance, SMETE is a learning object repository for the teaching and learning of science, mathematics, engineering and technology at all levels, that includes a recommender system based on past user interactions (Neven and Duval, 2002).

Several techniques are used for guidance and for providing recommendations to users. Among others, collaborative filtering (Herlocker, 2004) is one of the most successful ones. Briefly, collaborative filtering is selecting content based on the preferences of people with similar interests, basically by pooling and ranking informed opinions (or experiences of use) on any particular topic. That is to say, an automatic system collects information about user actions (explicit, such as voting or answering a question; or implicit, such as noticing which offered links are visited and which are not, how often and how much time is spent) and determines the relative importance of each content by weighting all the collected information among the large amount of users. Both navigational techniques are also valid in a digital library scenario. As stated in Fourier (2006), some authors found that personality types and learning styles will influence information-seeking styles (Limberg, 1999). Therefore, searching and browsing activities can be a useful source of information about user behavior.

Nevertheless, it seems clear that all the information needed cannot be stored in the learning objects in the form of metadata, as they would become too specific,

thus reducing reusability, the main goal of any learning object repository. The use of external structures for supporting these needs can be implemented by means of ontologies which describe the relationships between all the elements in the learning scenario. Ontologies can also be used to better describe such elements, incorporating a semantic level of information which can be used to enrich the learning process (Sicilia and García-Barriocanal, 2005).

The basic idea of this paper is that the interactions with learning objects carried out by different user profiles can be stored in a structured way, and then shared for future users with similar necessities, overcoming information overload and difficult quality assessment. Besides this, the analysis of such interactions may also reveal interesting facts about the use of learning resources that can be added as new metadata to such resources, improving their overall quality, but also in the learning process itself if it is described by means of formal descriptions, such as those provided by the IMS-LD standard, for example.

2.1 The UOC learning object repository

From a teacher's perspective, setting up a learning object repository and providing contents is not the most important issue in a learning scenario. On the contrary, the most interesting information is extracted from the usage that learners perform on such a repository. The cost of setting up a repository of learning objects needs to be justified, at least, by a high degree of use, and by a continuous feedback that allows teachers and instructional designers to extract useful information from the learning process followed by learners. Quality is also one of the major concerns for any repository, as learners will use it only if they feel confident of the available contents. Although quality is ensured by the institution setting up the repository, it is also important to allow users to participate in a continuous quality improvement process, both explicitly, by means of user ratings or annotations, or implicitly, by means of analyzing the usage of the repository and inferring relevant patterns.

In the particular case of the Universitat Oberta de Catalunya (UOC, in English known as Open University of Catalonia), a pure online university where students and teachers interact by means of a virtual campus, a shift from a content-based towards an activity-based learning process is now under development, although the pedagogical model was already designed under a user-centered approach. The UOC e-learning environment can be considered a lifelong learning scenario, where both contents and the learning process are provided by the institution, which ensures a high degree of quality control in a top-down approach.

There is evidence that teaching methods are shifting from a transmission of knowledge to a problem-based learning process. This increases the use of libraries, collections and repositories (Limberg, 1999). Following this quote, the UOC pedagogical model is based on a new model for teaching and learning where the teacher becomes a guide in the learning process, for which the student is ultimately responsible (Sangrà, 2003). For each course, the teacher establishes a learning plan where a calendar, an activity schedule, the basic

communication tools and a suggestion of learning objects available at the institutional repositories are proposed.

These learning objects are stored in the digital library framework in two different repositories, depending on its source type. First, the “OPAC” repository (Online Public Access Catalog), where the recommended bibliography is stored, which is linked to the service that provides digital versions of chapters of available books. The OPAC also includes the subject textbooks in HTML and/or PDF formats. Second, the “Digital Collection” repository, with content from external providers subscribed to by the academic library such as academic databases, electronic journals, as well as free Internet resources, proposed exercises and previous exams, or theses and dissertations done by teachers and students of the university from previous semesters. According to Fox and Shalini (2002), the UOC learning object repository is a client-server based approach, as opposed to peer-to-peer approaches, as a basic policy for ensuring quality issues in the learning process. Nevertheless, not all external resources used by learners are known, so it is important to understand that only partial knowledge about the information behavior is available.

The available contents are located at the virtual campus, either in the digital library as a whole, or as a subset placed in each virtual classroom. Students and teachers have free remote access to the digital library, but only the students registered to one particular course have access to its specific classroom library. In any case, it is also well know that students use external information sources for accomplishing their learning goals, and teachers might also recommend the use of Internet search engines for doing so. In order to know the real implications of this fact, in a survey performed by the digital library on students enrolled on the first semester of the 2004-05 course, 31% of the students stated that they start searching for information for educational subject purposes from the library resources, while 53% start straight ahead from an Internet portal or a search engine (UOC, 2005).

The learning objects stored at the library are catalogued in MARC 21 if they are accessible trough the OPAC, or in Dublin Core if they are accessible from the Digital Collection. Currently, there is an ongoing project at the university for cataloguing all the subject textbooks using the LOM standard, for satisfying the growing needs of the university, while the MPEG-7 standard is also under evaluation for description purposes (Pascual et al., 2006).

2.2 Access and navigational profiles

Preliminary studies have been carried out in order to obtain an initial picture of the interactions between learners and the virtual campus, including both services and resources usage. These studies try to establish relationships between user profiles and their information and navigational behavior in the virtual campus, with the aim of promoting personalization and improving the understanding of what learning in virtual environments means.

It is worth taking into account the particularities of UOC students. The most common profile is an adult with an average age between 26 and 35 years old

(57%), married and with children (55%), with a full time job (93%) and, a very important issue, who already has a previous university degree (60%), but wants to be updated and improve his or her knowledge, either for personal or professional reasons. A study of the satisfaction of graduate students shows that they chose the UOC for the learning model as the main reason because it is a fully distance online system that allows them to study from anywhere at anytime, and it is very flexible. The concept of lifelong learning is also important, as 38% of graduate students have also chosen the university because they wanted to improve their knowledge, and 44% of graduate students have chosen a degree related to their job because 28% of students wanted to improve in the exercise of their professions (UOC, 2005).

A first internal study (UOC, 2005) was carried out to determine the way learners interact with all the educational resources available through the virtual classroom, the digital library and the Internet. A total of 1108 students, covering 20 different degrees, were asked to participate in a survey where their behavior with respect to learning resources was analyzed. User profiling by means of segmentation was carried out using the following variables: how often and from where (and how) they access basic learning resources (R, including the subject textbook), the communication spaces in the virtual classroom (V), additional bibliography (B) and optional further resources and readings (F); how often they use these resources for preparing the continuous evaluation activities, and how often they use them for preparing the final validation test or exam. For example, there is a question related to place how and from where they access the resources, which combines the use of computers (D, in the form of HTML or PDF digital resources) or paper textbooks (T), from home (H), office (O) or public spaces (P, for example libraries or cyber cafés), or while commuting (C). Each possible combination of values is dichotomized in order to show whether a combination is present or not, converting categorical scales into binary variables, grouping contiguous values. A total of 11 binary variables were relevant for clustering purposes, using a non hierarchical typological analysis. User satisfaction with respect to the learning resources usage was evaluated according to this setup.

Four typologies were discovered, as shown in Table I. Capital letters are used to show a strong use or relationship, weak otherwise.

Typology	**Where**	**Format**	**Resources**	**Satisfaction**
Standard (54%)	H, O	T	R, V	Very high
Explorers (25%)	H	D, T	R, V, F	Very high
Involved (11%)	H, o	D, T	R, V, B, F	High
Non-involved (9%)	h	T	R	Medium / Low

Table I. Hierarchical typological analysis results.

In light of these results, it is clear that textbooks based on paper are still a very important element in the learning process, probably, as some authors have stated, because the development of e-books has been led primarily by technology instead of by users' requirements, and the gap between functionality and usability is sufficiently wide to justify the lack of success of the first

generation of e-books. Therefore, it is crucial to identify the needs and requirements of the target community so that the design can fulfill their needs and expectations. The acquisition of a well-defined user profile is an essential component of the design process for the successful development of e-books (Landoni and Diaz, 2003).

From the table shown above, we can see that half of the students (54%) use only the subject textbooks and read the messages posted in the classroom and that one third of students (36%) search and use additional educational resources such as bibliography and further readings. In order to prepare the assignments, all the students (95.5%) use the paper textbook which is sent to the students' home at the beginning of the semester. And only 8% of the students always use the bibliography accessible from the library classroom for preparing the assignments and only 10% the one recommended by teachers. On the other hand, half of the students say that they frequently (30.4%) or even always (19%) search the Internet to find documents that will help them in preparing assignments.

Other additional variables used in this study also showed interesting information that can be used for personalization purposes. For example, students in the "Involved" typology, where the oldest was around 42.5 years old, showing that the age group might be used as a relevant variable for selecting resource types, as a reasonable indicator of previous study habits.

Another interesting study (Carbó et al., 2005) related to navigational behavior was carried out with students from several subjects from the Computer Science degree. In this study, the main goal was to establish a relationship between navigational behavior and academic performance, according to each scheduling (which is different for each subject). Preliminary results show that there are three different navigational patterns: first, students that connect every day or almost every day; second, students that mostly connect on weekends (from Friday to Sunday, both inclusive); and third, students that only connect when they have to deliver an exercise or according to the published scheduling (i.e. to participate in a discussion in the virtual classroom forum). Navigational patterns and interaction levels are strongly related to academic results, and very simple rules can be extracted from the interactions during the first week of a scheduled exercise, for example. These experiments were designed taking into account mainly user interaction, but learning resources usage could also be another interesting parameter to include in such experiments, especially when learners are supposed to use specific resources provided by the teacher for solving an exercise.

Therefore, interactions must be studied at different levels, among all the different elements in the e-learning scenario, depending on the context being evaluated. When the use of learning resources is involved, it is important to gather information about which resources are accessed, when and, if possible, how they are evaluated by learners. This information could be used to improve resource visibility or ranking, according to previous experiences by similar users, for example.

3. Usage data harvesting and analysis

Our proposal consists of identifying and describing the elements in the learning scenario, namely users, resources and the learning process in itself, and establishing the relationships that occur between them during the basic interactions performed by users, extracting relevant information for our purposes.

In the case of the learning scenario of the Open University of Catalonia, students log in to the virtual campus and have access to several services. Among them, the mailbox service and the virtual classroom are the most important. In the virtual classroom (one for each subject a student is enrolled on), students find a teaching plan, a calendar, a set of learning resources and several notice boards and forums. Students are expected to follow the teaching plan according to the calendar, which guides them through all the learning activities they must perform, interacting with teachers and other students through the notice boards and forums, and using the selected learning resources but also other additional ones available through the digital library or external search engines. In an ideal scenario, the teaching plan is a dynamic learning process integrated in an intelligent tutoring system giving support to the virtual campus, providing students with adaptive learning itineraries (Mor and Minguillón, 2004).

Therefore, there are several types of interactions that might be relevant for analysis purposes in such a scenario. The elements of the e-learning environment are learners, teachers, learning resources available through a repository, services (i.e. the digital library), and the learning process in itself, which can be seen both as a complex dynamic service and as a special kind of resource as well. In this paper we are not interested in modeling interactions between students or between teachers and students, but those dealing with learning resources.

Following the ideas presented in Ferran et al. (2005), we propose to use an ontology for describing the learning process, as the core of the intelligent tutoring system. This ontology will use other sub-ontologies for handling all the interactions with the learning object repository, the interactions with other services and the user profiles. Each ontology is responsible for determining which information is relevant for usage analysis, and this information is shared among the different ontologies, making the learning management system (i.e. the virtual campus) aware of the interactions. For example, when students search for learning resources, the ontology responsible for the searching process in the repository uses information from the student profile, in order to select the most appropriate resources according to learning style and accessibility issues.

3.1 Building a user model

In order to build a multidimensional user model and feed it from usage data, several fundamental questions must be addressed, as stated in Smeaton and Callan (2005):

- What data should (and can) be collected and how can be captured?
- How are anomalous data recognized and filtered out?
- How should the data be analyzed and which parameters need to be set?
- How are data weighted appropriately over time?

The first question is partially answered through what is known as deep log analysis techniques (Nicholas et al, 2006). Basically, it consists of triangulating and enriching data from all possible sources, namely campus navigational logs, library usage logs, socio-demographic data and academic background. These data are captured in different ways. Both navigational and library usage data are stored in web servers as log files, usually following a standard such as the Apache Common Log Format (CLF). Socio-demographic data and academic background are provided by students during the enrollment process, and they are updated each semester. All this information should be stored using a standard format (IMS LIP, 2005), in order to promote sharing with other institutions and services.

The second question involves the preprocessing stage of the collected data. Several common problems must be addressed: first, log server files are huge, around 50GB each week, with millions of lines to be processed, although less than 1% of the lines contain useful information for navigational pattern analysis. Even on a daily or hourly basis, performing such an analysis may be computationally prohibitive. This can be partially solved by introducing specific marks in the web site and then filtering out those lines not containing these marks. This approach has two important advantages: first, the resulting log files are much smaller, and second, as marks are directly related to user actions, it is much easier to track users' real intentions. Other problems related to the use of log server files are the possible collisions for the IP addresses identifying each connection, because of the proxies used by the Internet providers. Nevertheless, this is not a real issue if users are uniquely identified when they log in to the web site, as is the case of the UOC virtual campus, where each user session is uniquely identified and, therefore, it can be tracked for analysis purposes. Once preprocessing is done, all available data must undergo a data mining process using the appropriate tools, using proprietary but also ad-hoc software, in order to mine the raw data more sophisticatedly. Clickstream techniques (Mobasher, 2002) try to discover navigational patterns that can be related to user tasks, by combining the identified actions in the preprocessed log files. On the other hand, simple statistical analysis can be also carried out to extract useful information, for example, which are the most common keywords used for locating a learning resource.

Finally, a particular challenge for personalization is that long-term models must encompass a time span that is defined in terms of a human lifetime (Gemmel et al, 2003; Smeaton and Callan, 2005). Furthermore, such models also need to incorporate the learning scenario scheduling, that is, the concept of academic semester in the case of higher education, for example. This is important

because user actions are determined by such scheduling. On the other hand, users in complex environments such as the UOC virtual campus receive multiple inputs from different spaces and services, showing different behaviors depending on many events and variables (scheduling, experience of use, and so on). Therefore, it is important to understand the real motivations that are the underlying cause for explaining user behavior. This can be partially accomplished by means of surveys and user tests, where quantitative, but especially qualitative, data about system usage are much better obtained. Nevertheless, this point is outside the scope of this paper, so it will not be developed here.

3.2 Experimental results

With the available data from one academic semester (from February to June 2006), an experiment has been performed in order to determine the usage of several learning resources available in the virtual classroom library space. We have chosen the user behavior data generated during the first assignment, where a few learning resources are supposed to be used by learners in order to solve the proposed learning activities. Our purpose was to determine the actions that the learners perform with the proposed learning resources in order to find information that will help in the learning design (LD) of the next semesters as well as for improving the description of the learning objects, both of them objectives for personalization purposes.

We have analyzed the data from the usage performed by students with the learning objects available in two different virtual classrooms from two subjects, Statistics from the Computer Science degree, and Managing Archives from the Information Science degree. The Statistics subject had 280 students, while Managing Archives had only 60 students. Both subjects are mandatory for students if they want to get their degree. Learners have access the resources through the virtual campus while they are in the virtual classroom. We do not analyze the specific usage of individual users, as we are interested in detecting typical behaviors. We use the log server files generated by the Apache web servers which act as front-ends for the virtual campus.

We will show a few examples of interesting student behaviors that we have detected that could help us to enrich metadata descriptions, even automatically, and we will also discuss the limitations of this method.

3.2.1 The Statistics subject

In this case the students have to solve several exercises about descriptive statistics. They have available several examples similar to these exercises, and they also have a document with all the errata in the textbook, which is supposed to be read before the exercises are solved. Students also have a guide for planning the work to do which helps them to establish an appropriate pace for learning all the concepts needed to solve the exercises. They have almost three weeks for reading the guide, the textbook (incorporating the changes described in the erratum file), and using similar exercises as learning examples before they try to solve the exercises.

The usage data captured during these three weeks shows that some of the available examples are never used. One possible reason is that those examples are poorly tagged and then become “invisible” to students. Another reason is that those examples are not relevant for solving the first learning activity. In any case, this information could be useful for narrowing the searches performed by students when they try to find examples for this first learning activity. A “relevance” factor could be added to each learning resource with respect to each learning activity, according to the gathered data.

On the other hand, the erratum file is downloaded by most users, as expected, together with the learning guide. Therefore, the intelligent tutoring system could use this fact to warn students who have not done so (and probably their teachers and tutors too) that the date for delivering the solution of the learning activity is near. This rule could be incorporated in the LD description of the learning activity and triggered by the ontology responsible of the learning process.

It is also interesting to note that the erratum file has a different attitude with respect to the other resources: it is downloaded by the students following an exponential decrease; it is downloaded mostly at the beginning of the semester, whereas the rest of resources, for instance the learning guide but especially the additional exercises and examples, are accessed following teachers’ recommendations or according to the teaching plan, as shown in Figure 1. Therefore, any usage analysis must be contextualized in the period of time where it is carried out, because different results can be obtained.

‘take in Figure 1’

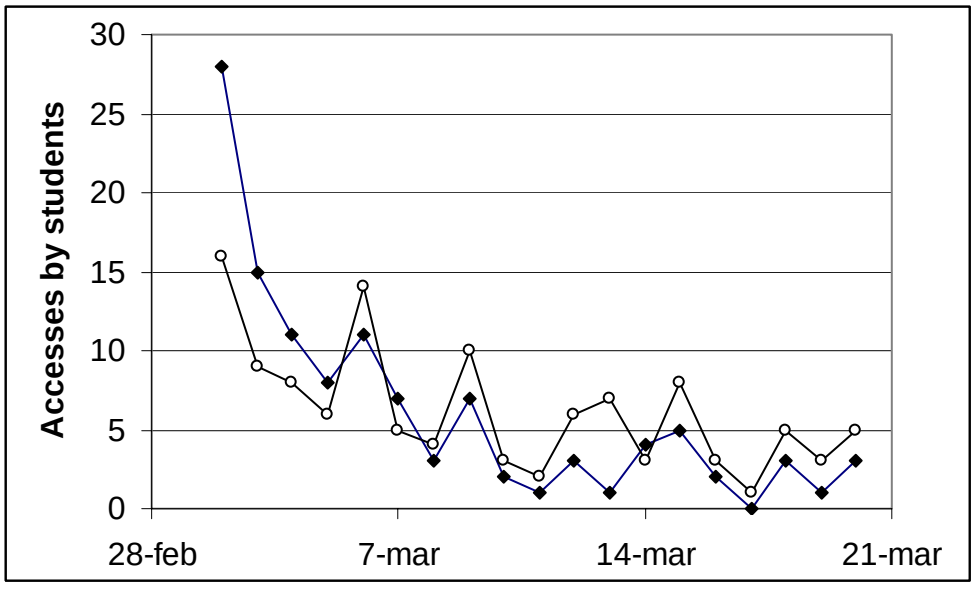

Figure 1. Dynamics of students’ accesses to different types of files. Closed symbols are the “Erratum file” and open symbols are the “Learning guide”.

3.2.2 The Managing Archives subject

On the other hand, in the subject “Managing Archives” we analyzed the behavior of students with another type of learning object, specifically exercises. Students had the statements in one file and the corresponding solutions in another available in the virtual campus.

The usage data confirmed that students downloaded the files in the order expected from the teaching plan. Indeed, in the first place students downloaded the statements, and after about two days, they downloaded the solutions (Figure 2).

‘take in Figure 2’

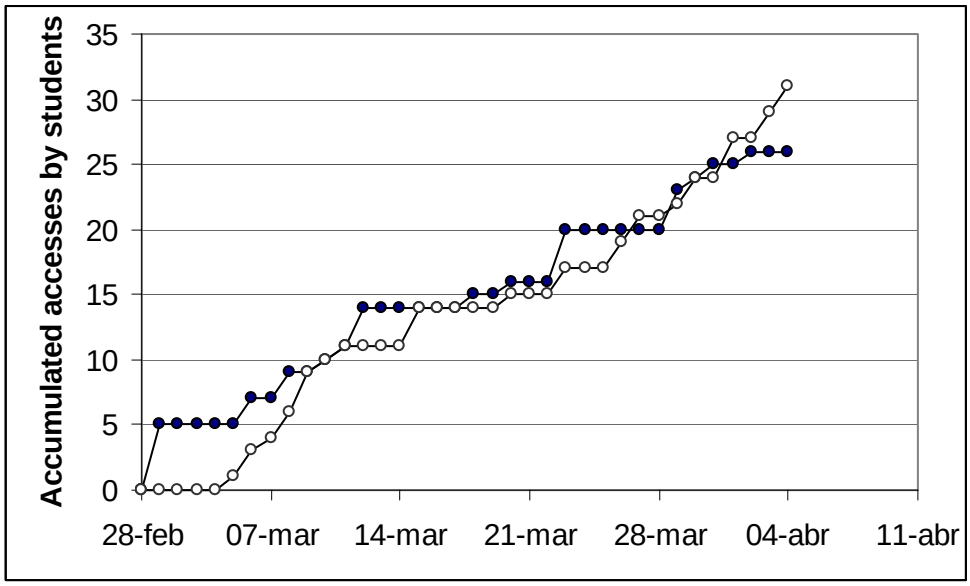

Figure 2. Dynamics of accumulated student’s accesses to the exercise files. Closed symbols are the “Statements” and open symbols are the “Solutions”.

This type of test confirms that students effectively follow the teaching plan and that they try to solve the problems themselves before checking the solutions. A departure from this pattern would alert of negative studying habits. Also, each exercise statement can be rated by parameters such as its usage and the delay for accessing its solution.

Another possible utility of monitoring student access to resources is to find and rate LO. For instance, while a student is navigating in the class he or she may access the library or the Internet in order to find complementary resources. Here, the resources accessed by students of a given subject inside the virtual library of the UOC were effectively detected in our experiments. Therefore, this can help teachers discover other materials that could be included in the virtual library space of the classroom or added to the teaching plan.

These were some simple examples of students’ behavior when interacting with learning objects that appear in the virtual classrooms. What is important to note is that these observations were derived from raw data logged by default by web servers – i.e. no special utilities were required in the virtual campus for

recording these raw data. This means that the same approach can be used in other learning environments as well. However, this approach works at an HTTP session level and it is not able to identify and track the activities of a given student performed in different sessions. The anonymity of students implies that their behavior in a session can neither be related to previous sessions nor to other data from that student, such as his/her academic success.

In order to get a more in-depth view of the information on student behavior, more detailed data are required and, hence, specific software intended for this purpose should be installed on the virtual campus. Further studies need to be carried out in order to have a complete view of the information behavior for e-learning purposes.

3.3 Enriching metadata in e-learning environments

Nowadays metadata used for describing the resources located in the virtual classrooms follow the Dublin Core initiative and are provided by librarians. When students retrieve these resources, they see title, author and other authoritative fields only.

Other metadata are used internally for librarians and lecturers for searching in the digital e-learning repository. Currently, the repository cannot be completely accessible and searchable by the whole student community due to copyright laws that establish some restrictions for some resources, and only students enrolled on a particular subject are allowed to access such resources. Therefore, resources provided by the university in the virtual classroom are usually in a directory format, so metadata are not used by students for discovering the materials, although authoritative metadata (Recker and Wiley, 2001) can be automatically produced. Following the standard for Learning Object Metadata base schema (LOM, 2002) we suggest how a selection of LOM fields are involved in a process of metadata enrichment with information extracted from usage:

- LOM 1.5. *Keyword*: keywords or phrases describing the topic of the learning object. Through the searches performed by librarians and, when the repository is accessible for teachers and students, also by their searches, it is possible to rank keywords according to the number of times they are used for retrieving the given learning object and, therefore, be used for improving recommendations or detecting misplaced keywords, for example.

- LOM 1.7. *Structure* and 1.8 *Aggregation level*: underlying organizational structure of the learning object. Usually, learning objects are described as independent chunks of information so they are considered to have an atomic structure, which is supposed to be indivisible. Usage data may reveal relationships with other learning objects and this fact can be used to create collections or hierarchical, linear or networked structures between them.

- LOM 3. *Metadata*: this field describes the metadata record itself.

It could be used for registering all the automatic changes that the system performs, in order to further analyze the metadata enrichment process itself.

- LOM 5. *Educational*: this category describes the key educational or pedagogic characteristics of the learning object. Currently, it is one of the most criticized aspects of learning objects and LOM, as it is clearly underused. It is related to the quality of the learning experience, so it becomes critical for any intelligent tutoring system dealing with learning objects for building adaptive itineraries.

    - LOM 5.1. *Interactivity type* and 5.3. *Interactivity level*: these fields can indicate active learning, expositive or mixed, and its degree. These fields can be linked with the information search characteristic registered in the user profile. We could see, for example, that some resources provided on the “Statistics” course (additional exercises and examples) were not used by all users, as some of them preferred to base their study on the use of the hypertext material, which is more theoretical and textual instead of practical. Other users prefer learning-by-doing instead and, therefore, under a personalized learning process they could be recommended to use learning objects with such an interactivity type and level.

    - LOM 5.8. *Difficulty*: how hard it is to work with or through this learning object for the typical target audience. In an explicit way, users (the teacher but also the students) could suggest values for this field and the system could use it for selecting exercises for students according to their profile.

    - LOM 5.9. *Typical learning time*: approximate or typical time it takes to work with or through this learning object for the typical intended target audience. For pure online learning objects (i.e. exercises with applets or simulations), the system can estimate the average time and use it to detect “outliers”, that is, people that just walk through or people that spend too much time, taking the appropriate actions in each case.

- LOM 7. *Relation*: this category defines the relationship between this learning object and other learning objects, if any. This category can be used by the intelligent tutoring system to establish relationships between learning objects according to their usage, especially those detected from the adaptive paths followed by students.

- LOM 9. *Classification*: this category describes where the learning object falls within a particular classification system. As an extension of the concept of keyword, it is possible to analyze the usage of the terms in the taxonomy for discovering interesting relationships and unused terms.

As regards the learning process itself, which is handled by the intelligent tutoring system and the associated ontology, it is interesting to establish appropriate relationships between the proposed adaptive paths, their degree of acceptance by learners and their academic performance, as a valuable feedback for teachers and instructional designers. Following the same approach, these relationships could be incorporated in the metadata describing such itineraries following the IMS-LD standard.

3.4 Balancing privacy issues and social effects

Assurance and trust are considered the most important drivers of *e-service* satisfaction and loyalty (Reichheld and Schefter, 2000). Assurance results from perceptions of security, safety and trust. Security is the extent to which customers perceive the provider's web services to be free from intrusions by third parties, whereas privacy refers to the active maintenance of a level of confidentiality with respect to private information provided to the provider. Users expect to be able to trust organizations to protect any personal information they may have gathered. Preserving privacy and anonymity is an aspect that has been very important for libraries, and for this reason, an extended best practice of Integrated Library Systems is to destroy patron-related data. For example, the library only keeps the link between a patron and a book while the book is out on loan, in order to protect the library's holdings. But, once a book is returned to the library, all the patron activity history is deleted. This can explain why recommender systems are not likely to be used in libraries (Lynch, 2001).

On the contrary, a very important aspect that cannot be ignored is the fact that, in an Internet learning environment, users are always under control, in the sense that all actions are monitored and registered. This might seem a very invasive setup which harms user privacy and, therefore, undesirable. Nevertheless, there are several notable facts that need to be clarified: a) users know in advance that, in a web-based environment, all actions are logged; b) the recommendation system must be designed in a non-intrusive manner and be user-friendly, including the possibility of disconnecting it or minimizing its participation in the browsing or searching activities; and c) the participation of individual users in the final recommendation system is completely anonymous. Finally, it is also important to note that the information collected is not meant for commercial purposes, and that the library (a non-profit organization) will use the data rationally and in a transparent way.

Furthermore, e-learning environments and digital libraries "can serve as meeting places where people can communicate with each other through the documents, annotations, and logs they make available to each other, and through the conversation and discussion around this shared information" (Smeaton and Callan, 2005). This information exchange can be made in an implicit or explicit manner after the user consents to offering his or her usage data to the rest of the community. As usual, a tradeoff between personalization and privacy must be achieved (Kasanoff, 2001) in order to ensure the desirable social effects and a win-to-win scenario.

4. Conclusions

It seems clear that learning object repositories will become a basic element of any learning environment, providing users with high quality contents, properly described and supported by means of metadata, taxonomies and ontologies. The integration of such repositories into the learning process is a key issue for ensuring a proper use, not just being a mere space in which to find educational resources. E-learning success is somehow determined by the satisfaction in the learning process achieved by each student, and this satisfaction is directly related to the degree of interaction with learning resources, with the teacher and the other students in the virtual classroom, and the flexibility of the learning process in itself.

Furthermore, the forthcoming implementation of the Bologna process gives more responsibility to learners, making them the center of any formative action, promoting personalization, in order to adapt the learning process to each user particularities, needs and preferences, shifting towards a lifelong learning scenario. With the description of the learning process using competences and activities instead of contents, repositories need to be rethought in order to incorporate this new paradigm. Lifelong learning scenarios are based on a heavy use of available learning resources, where learners decide which contents are relevant for their purposes and which are not, with the possible guidance of an intelligent tutoring system. Discovering successful paths is a key issue for teachers and instructional designers for creating and updating such educational actions.

Content usage analysis is, therefore, a very important tool for ensuring that all the learning resources in a learning process are properly used, satisfying the quality policies established by the institution, and providing system designers with the relevant information about the real use of the e-learning environment. We have described several experiments we have carried out with real usage data from an academic semester in two different subjects from two official degrees offered at the university, showing that even simple experiments reveal information about student behavior which can be incorporated into their learning profile but also into the metadata used for describing the learning resources and the learning process itself, in order to improve all the information handled by the intelligent tutoring system, the core of the personalization process. A proposal for enriching the LOM standard has also been described, showing the non-authoritative metadata fields that can be automatically generated from the analysis of the interactions between learners and resources.

Currently, the UOC virtual campus is undergoing a major technological change for incorporating the new e-learning standards, such as IMS-LD, for example, with the aim of providing the learning process with formal descriptions, helping teachers and students to achieve their goals through a personalized learning process. Capturing external searches using Google or any other search engine, but also external database providers is also necessary to obtain a comprehensive view of the information search behavior of students. In order to validate the proposal presented in this paper, we are setting up a repository for all the available resources in one of the subjects (specifically Statistics), which

will be supported by an ontology which will capture all the actions performed by learners and will use this information to update the metadata used to describe such resources. We are also developing the ontology giving support to the intelligent tutoring system which will provide students with adaptive learning paths, according to their learning style and the interactions with the available resources.

References

ADL (2004), *Sharable Content Object Reference Model (SCORM) 2004*, 3rd edition, available at: http://www.adlnet.gov/downloads/300.cfm (accessed 8 November 2006).

Bologna Declaration (1999), The European Higher Education Area. Joint Declaration of the European Ministers of Education. Convened in Bologna on the 19th of June 1999.

Carbó, J.M.; Mor, E.; Minguillón, J. (2005), "User navigational behavior in e-learning virtual environments", *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 243-249.

Ferran, N.; Mor, E.; Minguillón, J. (2005), "Towards personalization in digital libraries through ontologies", *Library Management Journal*, Vol. 25, No. 4/5, pp. 206-217.

Foster-Jones, J; Beazleigh, H. (2002), "Metadata in the changing learning environment: developing skills to achieve the blue skies", *Association for Learning Technology Journal*, Vol. 10, No. 1, pp. 52-59.

Fourie, I. (2006), "Learning from web information seeking studies: some suggestions for LIS practitioners", *The Electronic Library*, Vol 24, No. 1, pp. 20-37.

Fox, E. A.; Shalini, U. R. (2002), "Digital libraries", *Annual Review of Information Science and Technology*, Vol. 36, No. 1, pp. 502-589.

Friesen, N. (2002), "Semantic Interoperability, Communities of Practice and the CanCore Learning Object Metadata Profile". *Alternate Paper Tracks Proceedings of the 11th World Wide Web Conference (WWW 2002)*, Hawaii.

Gemmel, J.; Lueder, R.; Bell, G. (2003), "The MyLifeBits lifetime store (demo description)", *ACM Proceedings of the 10th International Conference on the World Wide Web*, Hong Kong, pp. 1-7, available at: http://research.microsoft.com/~jgemmell/pubs/ETP2003.pdf (accessed 1 November 2006).

Herlocker, J.L.; Konstan, J.A.; Terveen, L.G.; Riedl, J.T. (2004), "Evaluating collaborative filtering recommender systems", *ACM Transactions on Information Systems*, Vol. 22, No. 1, pp. 5-53.

IMS LD (2003), *IMS Learning Design specification*, available at: http://www.imsglobal.org/learningdesign/ (accessed 15 November 2006)

IMS LIP (2005), *IMS Learner Information Package Summary of Changes*, available at: http://www.imsglobal.org/profiles/lipv1p0p1/imslip_sumcv1p0p1.html (accessed 1 July 2006).

Kasanoff, B. (2001), *Making It Personal: How to Profit from Personalization without Invading Privacy*. Cambridge: Perseus Books Group.

Landoni, M; Diaz, P. (2003), "E-education: Design and Evaluation for Teaching and Learning", *Journal of Digital Information*, Vol. 3, No. 4, available at: http://jodi.ecs.soton.ac.uk/Articles/v03/i04/editorial/ (accessed 7 November 2006).

Lee, G. G.; Lin, H. (2005), "Customer perceptions of e-service quality in online shopping", *International Journal of Retail & Distribution Management*, Vol. 33, No. 2, pp. 161-176.

Limberg, L. (1999), "Experiencing information seeking and learning: a study of the interaction between two phenomena*", Information Research*, Vol. 5 No. 1., available at: http://informationr.net/ir/5-1/paper68.html (01 July 2006).

LOM (2000), *LOM (Learning Object Metadata) working draft v4.1*,available at: http://ltsc.ieee.org/doc/wg12/LOMv4.1htm (01 July 2006).

LOM (2002), *Draft standard for learning object metadata*, available at: http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf (07 November 2006).

Lynch, C.A. (2001), "Personalization and recommender systems in the larger context: New directions and research questions", Keynote paper presented at *Second DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries*, Dublin, Ireland, June 18-20, 2001, available at: http://www.ercim.org/publication/ws-proceedings/DelNoe02/CliffordLynchAbstract.pdf (01 July 2006).

Mobasher, B.; Dai, H.; Luo, T.; Nakagawa, M. (2002), "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization", *Data Mining and Knowledge Discovery*, Vol. 6, No. 1, pp. 61 – 82.

Mor, E.; Minguillón, J. (2004), "E-learning personalization based on itineraries and long-term navigational behavior". *Proceedings of the 13th International World Wide Web Conference*, New York, NY, USA, May 19 - 21, 2004, pp. 264-265.

Neven, F.; Duval, E. (2002), "Reusable learning objects: a survey of lom-based repositories", Proceedings of the *Tenth* ACM *international conference on*

Multimedia, pp. 291-294, available at: http://www.cs.kuleuven.ac.be/cwis/research/hmdb/publications/files/Lorsurvey.pdf (01 July 2006).

Nicholas, et al. (2006), "Scholarly journal usage: the results of deep log analysis", *Journal of Documentation*, Vol. 61, No. 2, pp. 248-280.

Pascual, M.; Ferran, N.; Minguillón, J. (2006), "Integration of multimedia content and e-learning resources in a digital library" *Proceedings of the SPIE, Internet Imaging VII*, Vol. 6061, pp. g1-g11.

Recker, M.M.; Wiley, D. A. (2001), "A non-authoritative educational metadata ontology for filtering and recommending learning objects", *Journal of Interactive Learning Environments*, Swets and Zeitlinger, The Netherlands.

Sangrà, A. (2003), "Universitat Oberta de Catalunya. A newly created institution" In D'Antoni, S. (Ed*.) The Virtual University. Models & Messages. Lessons from Case Studies*. Paris: International Institute for Educational Planning, Unesco, available at : http://www.unesco.org/iiep/virtualuniversity/files/uoc.pdf (01 July 2006).

Sicilia, M.A.; García-Barriocanal, E. (2005), "On the Convergence of Formal Ontologies and Standardized e-Learning", *Journal of Distance Education Technologies*, Vol. 3, No. 2, pp. 12-28.

Smeaton, A. F.; Callan, J. (2005), "Personalisation and recommender systems in digital libraries", *International Journal on Digital Libraries*, Vol. 5, No. 4, pp. 299-308.

Towle, B.; Halm, M. (2005), "Design Adaptive Learning Environments with Learning Design", In Koper, R. and Tattersall, C. eds. *Learning Design. A Handbook on Modeling and Delivering Networked Education and Training*, Springer, The Netherlands, pp. 215-226.

UOC (2005), *User profile surveys*, Internal Technical report.

Velterop. J. (2004), "The myth of 'unsustainable' Open Access journals", Nature, available at http://www.nature.com/nature/focus/accessdebate/10.html (accessed July 2006).

Voss, C. (2003), "Rethinking paradigms of service: Service in a virtual environment", *International Journal of Operations and Production Management*, Vol. 23, No. 1, pp. 88-104.

Núria Ferran Ferrer, lecturer at the Open university of Catalonia (UOC) and associated lecturer at the Universitat Autònoma de Barcelona (UAB) at the department of Information Sciences and Communication. Courses conducted: Digital libraries, Information sources applied to several areas, Information systems. Degrees in Journalism at UAB (1998) and in Documentation at UOC (2003). Master in Information and Knowledge Society, IN3-UOC (2005). PhD student at UOC in information behavior studies in e-learning students.

Jaume Casadesús, researcher on sensor systems for data aquisition at IRTA. PhD in Biology at Universitat de Barcelona (1995), Bachelor on Computer Systems Engineering for UOC (2004), Master in Software Engineering at Universitat Politecnica de Catalunya (2006), and student of Computer Engineering at UOC.

Monika Krakowska, assistant with PhD at the Jagiellonian University in the Institute of Information and Library Science. Research in Information and Communication Technology, Information Literacy, Multimedia, New tools in Information Science. Socrates Programme coordinator in the Institute of ILS JU. 2006/2007 academic year - courses conducted in: Information sources, Multimedia in social communication, European Union information, Basis of the information and library science. PhD degree in Bibliology, Information and Library Science (2006).

Julià Minguillón (Barcelona, Spain, 1968) received his Ph.D. degree from the Universidad Autònoma de Barcelona (UAB) in September 2002, with a thesis work about machine learning by cascading limited depth decision trees. In January 2001 he joined the Universitat Oberta de Catalunya (UOC) where he is a faculty member. Since November 2006, he is the Vicedirector of the Internet Interdisciplinary Institute (IN3). He has developed learning resources for object oriented programming, abstract data types engineering and compiler construction. He is also involved in the integration of e-learning standards into virtual e-learning environments, such as LOM, SCORM and IMS-LD. His main research interests include the formal description of the learning process by means of ontologies, personalizing the learning process by means of adaptive itineraries based on reusable learning objects, and user modeling in virtual e-learning environments applying web mining techniques. He leads the PERSONAL project, the framework that articulates all the aforementioned research lines. He is also in charge of the UOC participation in the OLCOS (Open Learning Content Observatory Services) EU funded project.